**Homework 11. Rao-Scott Chi-square test for survey data**

**MSDS 6370**

**Submitted by: Hari Narayan Sanadhya**

1. Consider the data in the chart below. Suppose you were told that the data was collected by randomly sampling individuals.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Married** | **Previously married** | **Never**  **married** | **TOTAL** |
| BAD HEALTH | 107 | 87 | 5 | 199 |
| NO BAD HEALTH | 835 | 460 | 33 | 1328 |
| TOTAL | 942 | 547 | 38 | 1527 |

a. Carry out a Chi-square test of the following hypothesis. Do the calculations manually and using SAS.

* *H*0: Self-reported bad health is independent of marital status.
* *H*a: Self-reported bad health is not independent of marital status.

**SAS Code**

\* Load the dataset;

libname x1 XLSX '/home/harisanadhya0/sasuser.v94/MSDS 6370/Unit 11/Dataset.xlsx';

data hisp;

set x1.data;

run;

\* Verify if the import was successful;

proc print data=hisp(obs=10); run;

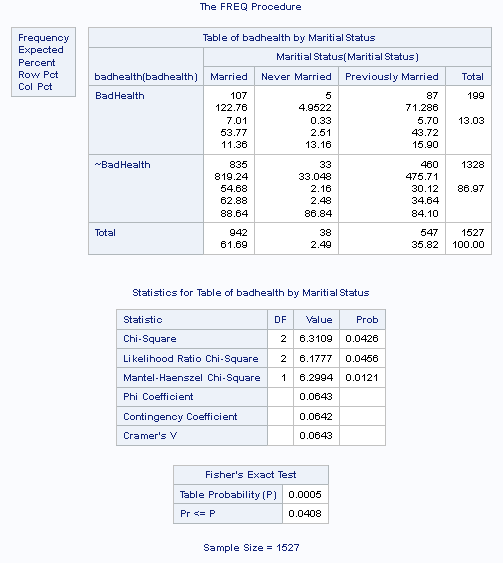
\* Generate the Chi-square statistics;

proc freq data = hisp;

tables badhealth \* MaritialStatus/exact chisq expected;

run;







DF 2

Alpha 0.05

Chi Square Critical Value 5.991

Since 6.31089 is greater then 5.991 i.e. Chi square value calculated is greater then the critical Chi-square value, reject the null hypothesis (df = 2, alpha = 0.05)



b. Write a sentence interpreting your result.

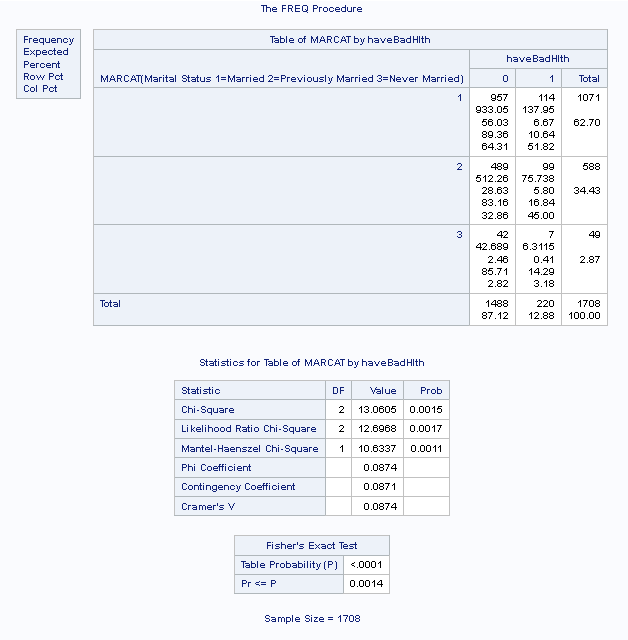
There is sufficient statistical evidence that the Self-reported bad health is not independent of marital status (based on Chi-sqaure test with DF=2 and 95% confidence). A greater proportion of Married individuals reported bad health as compared to Previously Married or Unmarried. Also a greater number of Previously married individuals reported bad health as compared to Unmarried. Also a greater number of Previously married individuals reported bad health then expected under the hypothesis of independence.

2. Now a colleague of yours familiar with the data set from which the values in the table above come from tells you that the data was collected with a sampling design containing clusters and strata. Using the data in the file sasDat.blt11\_4, reanalyze the data using the Rao-Scott Chi-square test to determine if marital status and self-reported bad health are independent or not.

Include your SAS code and output, as well as a sentence interpreting your output.

Note, the marital status variable in the data set is called marcat and you will have to use the code below to subset the data to Hispanics and create the variable haveBadHlth. Use proc contents to see all the variables in the data set that you will need to answer the question.

libname sasDat '/folders/myfolders/smuData/';  
  
data datin; set sasDat.blt11\_4;   
 if racecat = 1;  
 if kc001 = 5 then haveBadHlth = 1;  
 if kc001 NE 5 then haveBadHlth = 0;  
run;



SAS Code

libname x1 '/home/harisanadhya0/sasuser.v94/MSDS 6370/Unit 12';

data blt11;

set x1.blt11\_4;

run;

/\* Get the total number of records in the dataset \*/

proc summary data=blt11 print;

run;

/\* Above code gives output that the original set contained 18467 observations \*/

/\* Keep only Hispanics data and if kc001 = 5, then mark it as the person reported bad health otherwise not \*/

data blt11;

set x1.blt11\_4;

if racecat = 1;

if kc001 = 5 then haveBadHlth = 1;

if kc001 NE 5 then haveBadHlth = 0;

run;

/\* Print 10 observations to verify the data import \*/

proc print data=blt11(obs=10); run;

/\* Get the contents of the dataset \*/

proc contents data=blt11 varnum ;

run;

/\* Proc contents output showed that of the 18467 records, we have only 1708 observations left \*/

/\* Run the Rao-Scott Chi-square test \*/

proc freq data = blt11;

tables MARCAT \* haveBadHlth/exact chisq expected;

run;

There is sufficient statistical evidence that for Hispanic people, the Self-reported bad health is not independent of marital status (p-value of 0.0015 based on Chi-sqaure test with DF=2 and 95% confidence). The number of unmarried individuals that reported bad health were almost the same as what was expected under the hypothesis of independence but higher difference between expected and observed were in case of married and previously married individuals. More number of previously married individuals reported bad health then expected under the hypothesis of independence. 14.29% of the unmarried individuals reported bad health whereas only 5.8% of previously married individuals reported bad health and 6.67% of married individuals reported bad health. So more percent of unmarried individuals reported bad health when compared with married in case of Hispanic people.